Practice Problem

Given the weather conditions, we want to predict if a person is going for a run or not. The data that we have collected are the following:

	Sample	Features		Outcome
		Outlook	Wind	Run
Train	S1	Sunny	Weak	No
	S2	Sunny	Strong	No
	S3	Overcast	Weak	Yes
	S4	Rain	Weak	Yes
	S5	Rain	Weak	Yes
	S6	Rain	Strong	No
	S7	Overcast	Strong	Yes

Based on the above data, we will build a decision tree using the entropy splitting criterion. The input features are **Outlook** and **Wind**, while the outcome variable is **Run**.

(a) Compute the entropy splitting criterion of the outcome **Run** conditioned on the **Outlook** and **Wind** features. Which feature will be used as the splitting attribute in root of the tree? Show all your calculations.

Note: You **do not** need to perform arithmetic calculations for logarithms, e.g. if one of your equations contains $\log(\frac{1}{3})$, you can leave it like that and still solve the problem.

$$H(\text{Run} \mid \text{Outlook} = \text{Sunny}) = -\left[\frac{2}{0+2}\log\left(\frac{2}{0+2}\right) + \frac{0}{0+2}\log\left(\frac{0}{0+2}\right)\right] = 0$$

$$H(\text{Run} \mid \text{Outlook} = \text{Overcast}) = -\left[\frac{0}{0+2}\log\left(\frac{0}{0+2}\right) + \frac{2}{0+2}\log\left(\frac{2}{0+2}\right)\right] = 0$$

$$H(\text{Run} \mid \text{Outlook} = \text{Rain}) = -\left[\frac{1}{1+2}\log\left(\frac{1}{1+2}\right) + \frac{2}{1+2}\log\left(\frac{2}{1+2}\right)\right] = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}$$

$$H(\text{Run} \mid \text{Wind} = \text{Weak}) = -\left[\frac{1}{1+3}\log\left(\frac{1}{1+3}\right) + \frac{3}{1+3}\log\left(\frac{3}{1+3}\right)\right] = -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4}$$

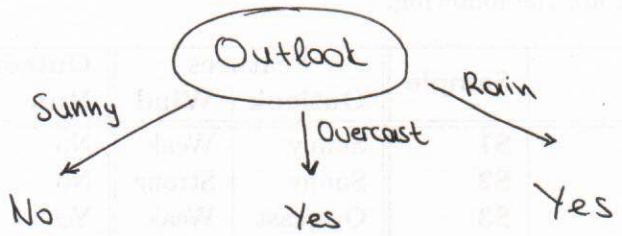
$$H(\text{Run} \mid \text{Wind} = \text{Strong}) = -\left[\frac{2}{1+2}\log\left(\frac{2}{1+2}\right) + \frac{1}{1+2}\log\left(\frac{1}{1+2}\right)\right] = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}$$

$$H(\text{Run} \mid \text{Outlook}) = \frac{2}{7} \times 0 + \frac{2}{7} \times 0 + \frac{3}{7}\left(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}\right)$$

$$H(\text{Run} \mid \text{Wind}) = \frac{4}{7}\left(-\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4}\right) + \frac{3}{7}\left(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}\right)$$

$H(\text{Run} \mid \text{Outlook}) < H(\text{Run} \mid \text{Wind})$ therefore Outlook is the splitting criterion for the root of the tree.

(b) Create the decision tree using only one node, i.e., the tree will only have the root. Please show the **splitting criterion of the node**, as well as **the decisions from each possible outcome of the corresponding criterion**. Please **describe how the decisions were made.**



All samples with "Sunny" correspond to "No" outcome.

All samples with "Overcast" correspond to "Yes" outcome.

The majority of samples with "Rain" correspond to "Yes" outcome.

(c) Which of the training samples will be classified correctly only using the above tree and which not?

Correctly classified: S1, S2, S3, S4, S5, S7

Incorrectly classified: S6

If I wanted to classify all samples correctly, I would have to extend the right branch of the tree based on samples S4, S5, S6

$H(Run|Wind=Weak) = H(Run|Wind=Strong) = 0 \rightarrow H(Run|Wind) = 0$

$H(Run|Outlook=Rain) = -\left(\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3}\right) \rightarrow H(Run|Outlook) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}$

$H(Run|Wind) < H(Run|Outlook)$ therefore splitting criterion is Wind